A Summary of Data Collection Procedures and Reports of Voter Turnout from the Current Population Survey

Kurt Bauman and Tiffany Julian
U.S. Census Bureau
Housing and Household Economic Statistics Division

November 1, 2010

Presentation to the Workshop on Overreporting of Voter Turnout
McGill University, Montreal Canada
November 5, 2010

A Summary of Data Collection Procedures and Reports of Voter Turnout from the Current Population Survey

Kurt Bauman and Tiffany Julian
U.S. Census Bureau

INTRODUCTION

The Current Population Survey (CPS), sponsored jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics, is the source for numerous high-profile economic and labor force statistics for the population of the United States, including the national unemployment rate. In 1964 and every 2 years since, a November supplement with questions on voting and registration has been administered at the end of the basic questionnaire. The resulting estimates of voting turnout have been widely used by researchers and the general public. For a history of the voting supplement, see Jennings (1990a).

Most estimates of voter turnout from the CPS and other sample surveys yield higher numbers of voters than reported by elections administrators (Jennings 1990b, Clausen 1968, Granberg and Holmberg 1991). On the other hand, the degree to which CPS differs from administrative records has historically been smaller than with some other surveys (e.g., Clausen 1968). Moreover, voter characteristics (age, sex, race) measured by CPS also match well with those recorded by election administrators (McDonald 2007).

This combination of circumstances leaves us in the position of facing two questions: (1) Why is voting overestimated in CPS relative to administrative sources? and (2) Why are CPS estimates often closer to administrative estimates than those of other surveys?

There are several possible explanations for the discrepancy between administrative and survey-based estimates of voter turnout. The first explanation is that the discrepancy is due to

problems with administrative records rather than shortcomings of survey-based estimates (Abramson and Claggett 1992, Cassel 2004). Since different survey estimates often provide significantly different estimates of turnout, this can only be a partial explanation.

The second explanation is that the problem lies in the process of asking people for information on voting. Social desirability bias may lead to overreport of voting among many respondents (Clausen 1968; Belli, Traugott, Young, and McGonagle 1999; Duff, Hanmer, Park, and White 2007). This may combine with various types of memory failure to increase overreport, especially as time elapses between the time of voting and the time of data collection (Belli, Traugott, Young, and McGonagle 1999). Similarly, surveys structured as a panel or that heighten awareness of voting might increase turnout estimates by stimulating overreport or by inducing respondents to go to the polls at higher rates (Presser and Traugott 1992).

A third explanation for higher estimates of voter turnout in sample surveys has to do with survey undercoverage and non-response. When sample responses are weighted to the full population, the implicit assumption is that those who could not be reached due to sample limitations have the same propensity to vote as respondents, which is probably not the case when measuring voting behavior (Burden 2000; McDonald 2003).

The CPS has strengths and weaknesses that might help explain why it tends to provide estimates that come closer to administrative reports than other surveys do, while also helping to explain why it differs from administrative estimates. What follows is a review of the CPS data collection and estimation process, with an emphasis on aspects that might be relevant to this difference. This will be followed by a direct comparison of estimates from CPS with state and national vote totals reported to the Clerk of the U.S. House of Representatives. The subsequent

3

section briefly examines the issue of survey undercoverage and non-response, and its effects on estimates. The final section offers conclusions.

ASPECTS OF THE CURRENT POPULATION SURVEY AFFECTING QUALITY OF VOTING ESTIMATES

This section of the paper will review procedures for collecting and reporting CPS data. Four topics are addressed: sample construction, questionnaire content, data collection and editing and processing.

Sample Construction

The CPS is administered by the Census Bureau using a probability selected sample of about 60,000 occupied households. The fieldwork is usually conducted during the calendar week that includes the 19th of each month, every month of the year. Households from all 50 states and the District of Columbia are in the survey for 4 consecutive months, out for the next 8 months, and then return for another 4 months before leaving the sample permanently. This design ensures a high degree of continuity from one month to the next (as well as between years). The 4-8-4 sampling scheme has the added benefit of allowing the constant replenishment of the sample without excessive burden to respondents.

A particular strength of the CPS is the effort that goes into constructing the sample. The sampling frame for the CPS starts from address lists from the fielding of the decennial Census. The address list is updated continuously for new housing built after the Census. Each decade a new sample is drawn, and then phased in over a period of months. Independent samples are

drawn from each of the 50 states, resulting in 824 sampling strata chosen for uniformity in terms of male and female unemployment rates, female headed households and households with 3 or more members. This process ensures that the CPS is representative of the broad range of household types and locations in the U.S. and improves the reliability of estimates.

Survey coverage is important for obtaining accurate estimates of voter turnout. The resources that are put into CPS sample construction may help explain why CPS estimates come closer than other surveys to estimates from administrative sources.

Questionnaire

The question on voting in the CPS is very similar to ones that have appeared in other surveys such as the American National Election Study (NES). The CPS question reads as follows:

> In any election, some people are not able to vote because they are sick or busy or have some other reason, and others do not want to vote. Did (you/name) vote in the election held on (date)?

Work on the NES has shown that question wording can affect response patterns (Belli Traugott Young and McGonagle 1999, Duff Hanmer Park and White 2007). In 2000, the NES question was changed from one similar to the CPS question to one that successfully reduced overreporting (see Appendix A), but the rate in that survey still remains higher than in CPS.

Question order can also influence survey responses, and it has been hypothesized that one

reason for overreporting of voting in surveys that focus on politics and voting is heightened awareness of these issues elicited among respondents (Clausen 1968). In contrast to some surveys of voting behavior, no questions on voting or politics precede the CPS question on voting. The question is followed by 3 or 4 questions (depending on response pattern) on how and when voting and registration occurred.   Depending on how many months the respondent has been in the survey, preceding questions focused on income, disability, earnings, or labor force status.

The question wording is probably not responsible for any lower level of voting report. However, the lack of other questions on voting and politics before the turnout question might contribute to lower estimates of voter turnout than NES and other surveys.

Data Collection

The Census Bureau invests considerable resources in training and preparation of interviewers in advance of data collection. This training, and the fact that the CPS is an official government survey result in response rates that are higher than those of many other household surveys. In November 2008, the household response rate was 91.2 percent. The supplement response rate was 91.8 percent, which, combined with responses of "don't know" or "refused," produce an overall response rate for the voting question of 87.3 percent.

An important aspect of the data collection process is the reliance on a single household respondent to collect information on other household members. This results in around 40 percent of voting responses obtained by proxy. As demonstrated by Highton (2005), proxy reports of voting tend to be lower than those of self-reports. Thus, the high rate of proxy

responses may contribute somewhat to the lower estimates of voting in CPS than in other surveys.

Although most CPS data are collected through phone interviews, the majority of respondents are contacted in person in the first month of the interview cycle, and again when they re-enter the sample in month 5 (after not having been contacted for 8 months). Overall, 34 percent of interviews are in person with the balance being telephone interviews. While it is unknown how survey mode affects response on voting questions, it is generally assumed that in-person contact improves the accuracy of survey response.

The CPS is designed for creating cross-sectional estimates of population characteristics, but the concern for continuity across months has led to the development of the 4-8-4 interview scheme. This gives the CPS some characteristics of a panel survey despite its cross-sectional nature. Households differ slightly in response rates and response patterns as their time in sample increases. We will see some evidence on this point later in the paper.

Overall, the data collection process probably serves to provide more conservative estimates of voting than may be provided by other surveys. High response rates and in-person visits likely improve the quality of estimates and may lower overreport. The use of proxy responses does not improve quality, but nonetheless lowers overreport. A question remains about the effect of time-in-sample.

Editing and processing

Data editing and processing procedures in CPS are similar to those in many other sample survey operations. Voting responses are collected electronically and relayed to Census Bureau

7

headquarters on a flow basis. The major step involved in editing most items in the data file is the imputation of missing data, generally by a hotdeck method.[1] Voting responses, however, are handled differently. Both historically and today, voting responses have been edited for consistency but not imputed.

With the lack of imputation, the procedures for estimating the number and percentage of the population that votes is different than for many estimates from the CPS. The number who report that they voted is taken as the total number who voted (once weights are applied), with no allowance for the possibility that some of those without a recorded answer might have voted. Because overreporting was a known problem, analysts felt that imputing from biased data would compound the problem of overestimation. The decision was made to estimate only the number of people who reported "yes," but make missing data available to those who might choose another method of estimating. These missing data are not assigned a value of "did not vote."

Weighting of the estimates corrects for certain kinds of missing data, but not all. In the weighting process, a non-response adjustment is applied. This corrects for what is termed "household non-response," which includes occupied households that refused to be interviewed, were temporarily absent, or otherwise could not be reached. A further weighting adjustment is made later in the process to bring estimated totals by demographic characteristic (i.e. black, female, total population) to levels determined by the population estimates program of the Census Bureau. This corrects for survey under-coverage. The remaining portion of non-response consists of people who were in a household that responded to the basic labor force portion of the

---

[1] The "hotdeck" imputation method fills the missing value with the response of another respondent with complete information, matched by demographic characteristics and geography. The result is an unbiased estimate that retains some of the variance of the complete set of responses.

survey, but did not answer the supplement questions on voting. In 2008, this came to 12.7 percent of respondents in interviewed households.

Table 2 shows a hypothetical population of 100,000 people in universe for the voting question (citizens 18 and over). Applying the coverage ratio from the 2008 CPS, 87,000 people would be in the sample of occupied households. With a household response rate of 91 percent (and assuming the number of people is proportional to the number of households), the number of people interviewed for the CPS would be 79,581. Only 69,506 of the people in these households would actually provide an answer to the question on voting. The other 10,075 (12.7 percent) have missing data on this item, either because they responded only to the labor force questions (supplement non-response) or because they provided and answer of "don't know" or "refused" to the voting question itself.

The weights are designed to make the population in interviewed households reflect the total population. The average weight will be related to the ratio of the two, that is, $\frac{100,000}{79,581}$ or 1.257. Since those who answered "yes" are taken to be the number who voted, they are simply multiplied up by the weight of 1.257 to equal 64,936. As a result, the proportion of the total population that is estimated to have voted is equal to the voting rate based on the interviewed population, rather than the voting rate based on respondents to the voting question.

The above estimate assumes the 10,075 people with labor force responses but no voting response did not vote. If they voted at the same rate as the people who did respond, the estimated turnout would be 74,349, or 9.4 percentage points higher. This is an approximation based on simplified assumptions.

We simulated imputation of 2008 voting data to estimate what the voting numbers might look like if missing data were to be imputed using a hot-deck method. This method used sex, education, age, and race to impute values for all cases that were in universe but did not have a value of "yes" or "no"—including "No Response", "Don't Know", and "Refused." The result was 20.1 million in the "yes" category and 8.2 million in the "no" category, moving 9.8 percentage points to the "yes" category[2].

The point of this demonstration is to show that the assumptions made during the editing and weighting process have very large impacts on the estimated voter turnout. The assumptions may not be particularly realistic – household non-respondents voted in the same way as respondents, supplement non-respondents didn't vote – but more realistic assumptions aren't readily available. More than any of the other factors examined so far, this is likely to have an impact on CPS overreport of voting.

COMPARISON OF VOTING ESTIMATES FROM THE CURRENT POPULATION SURVEY TO THOSE FROM ADMINISTRATIVE SOURCES

In this section, we will consider the performance of CPS voter turnout estimates across time and geography in relation to estimates obtained from the Clerk of the U.S. House of Representatives. Over the past four Presidential elections, the CPS turnout estimate was not statistically different from the "official" turnout estimate only in 2008 (see Table 1). In prior elections, the CPS turnout was higher by 3 to 8 million (3 to 9 percent higher than the administrative estimate).

_____

[2] See Appendix table 1 for more information

The trend in overreport from 1980 to 2008 is illustrated in Figure 1. In 1980, the CPS estimate was 7.5 percent higher than the administrative estimate, climbed to 11.6 percent higher in 1988, and decreased gradually to -0.2 percent in 2008.

The fact that overreport varies over time implies that an unchanging feature such as question wording is not the only factor at play. Overreport due to social desirability bias might shift over time with turnout due to social pressure to report voting (Bernstein, Chadha and Montjoy 2001). Conversely, overreport due to social desirability might shift in the opposite direction from turnout as a larger number of non-voters have the possibility of falsely reporting a vote (Anderson and Silver 1986; McDonald 2007). When we compare the pattern of turnout in Figure 2 to the overreport patterns in Figure 1, no obvious support for either theory exists.

Time in sample effects from interviews of people who have participated in the CPS interview in previous months may also affect overreport, but the basic 4-8-4 interview pattern has remained steady over the period examined in Figures 1 and 2. Figure 3 provides additional evidence on this point. The pattern of non-response mirrors the proportion who answered "yes" to the voting question in the first months in sample and then evens out it the later months.

Overreport also varies by state. The map in figure 4 shows the pattern of overreport appears not to be random. Several of the states with higher rates of overreport in CPS are found in the South and Southwest. Some of the states with low overreport are in the Midwest, the Mountain states and the Northeast.

Table 3 shows the patterns of state overreport across the three Presidential elections. Although there is shifting in the ranks from year to year, a number of states such as Colorado, South Carolina and Mississippi remained in the same quintile all three years. Over half of all

states remained in two adjacent quintiles across all three years. This reinforces the idea that there may be something systematic about the factors that lead to state-level overreport in CPS. It's beyond the scope of this paper to examine this, but it appears to be a promising area for further research.

The only remaining issue that can be addressed with a simple data examination is the effect of response rate and missing data. Figure 5 shows the trend of household-level missing data (the type addressed by weighting, note though that this excludes undercoverage, another factor in weighting) across a large span of the history of CPS. The trend is clearly upward, with an increase in household refusals since 1990. There is nothing in this pattern that suggests the overreport shown in Figure 1. Missing data on the voting question (the type that is not accounted for in weights) from 1996 to 2008 are shown in Figure 6. The increase in non-response corresponds to a decrease in overreporting during these years (Figure 1.) However, the time series is too short for us to be able to draw any conclusions. If we were to combine both types of non-response (household non-response and question non-response) an altogether different pattern might emerge. Once again, such an exploration is outside the scope of this paper, but appears to be worth pursuing.

DISCUSSION AND CONCLUSION

This paper has raised a number of possible explanations for the particular pattern of voting overreport found in the CPS. Question wording, question order, the use of personal interview, the use of proxies and the panel structure of CPS are things that have changed little, and thus do not provide a sufficient explanation for the changing levels of overreport over time.

The panel structure of CPS, in particular, seems to have a small impact if any. This does not imply that question wording and other factors make no difference. Prior research suggests strongly that question wording and the use of proxies do matter. The effect of unchanging aspects of the CPS may be to produce or increase overreport while other processes change the level of overreport across time and across geography.

Survey coverage, non-response and weighting may be a part of the process that produces changing overreport. CPS has high coverage and low non-response compared to many other household surveys. Nevertheless, the levels that remain have a large impact on estimates of voter turnout. The standard assumption that non-respondents resemble respondents may not hold in the case of voting reports. In practice, estimates from the CPS have abandoned this assumption by treating non-respondents to the voting question as if they haven't voted.

That the reasons for overreporting of voting in surveys has been examined for decades without resolution shows how difficult it is to find the answer to basic measurement issues. As this topic is explored in the future, continued examination of the role of survey non-response, at least in the case of CPS, is clearly warranted.

References:

Anderson, Barbara A. and Brian D. Silver. 1986. "Measurement and Mismeasurement of the Validity of the Self-reported Vote." *American Journal of Political Science* 80:771-785.

Abramson, Paul R. and William Claggett. 1992. "The Quality of Record Keeping and Racial Differences in Validated Turnout." *Journal of Politics* 54(3):871-80.

Belli, Robert F., Michael W. Traugott, Margaret Young and Katherine A. McGonagle. 1999. "Reducing Vote Overreporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring." *Public Opinion Quarterly* 63(1):90-108.

Burden, Barry C. 2000. "Voter Turnout and the National Election Studies." *Political Analysis* 8:389-98.

Cassel, Carol A. 2004. "Voting Records and Validated Voting Studies." *Public Opinion Quarterly* 68(1):102-08.

Clausen, Aage R. 1968. "Response Validity: Vote Report." *Public Opinion Quarterly.* 32:588-606.

Duff, Brian, Michael J. Hanmer, Won-Ho Park and Ismail K. White. 2007. "Good Excuses: Understanding Who Votes with an Improved Turnout Question." *Public Opinion Quarterly* 71(1):67-90.

Granberg, Donald and Soren Holmberg. 1991. "Self-Reported Turnout and Voter Validation." *American Journal of Political Science.* 35(2):448-59.

Highton, Benjamin. 2005. "Self-Reported Versus Proxy-Reported Voter Turnout in the Current Population Survey." *Public Opinion Quarterly* 69(1):113-123.

Jennings, Jerry T. 1990a. "The Current Population Survey of Voting and Registration: Summary and History." In, *Studies in the Measurement of Voter Turnout.* Current Population Reports, Special Studies, P23-168. Washington, DC: U.S. Census Bureau.

Jennings, Jerry T. 1990b. "Estimating Voter Turnout in the Current Population Survey." In, *Studies in the Measurement of Voter Turnout.* Current Population Reports, Special Studies, P23-168. Washington, DC: U.S. Census Bureau.

McDonald, Michael P. 2003. "On the Overreport Bias of the National Election Study Turnout Rate." *Political Analysis* 11:180-86.

McDonald, Michael P. 2007. "The True Electorate: A Cross-Validation of Voter Registration

Files and Election Survey Demographics." *Public Opinion Quarterly* 71(4):588-602.

Presser, Stanley and Michael Traugott 1992. "Little While Lies and Social Science Models: Correlated Response Errors in a Panel Study of Voting." *Public Opinion Quarterly* 56:77-86.

U.S. Census Bureau. 2006. *Current Population Survey Design and Methodology.* Technical Paper 66. Washington, DC: U.S. Census Bureau.

Table 1. Presidential Turnout 1996 to 2008

(In thousands)

Election Year	Total Population	Total Citizen Population	Reported Voted[1]	Margin of Error	Percent of Total Population	Percent of Citizen Population	Administrative Records[2]	Difference
2008	225,499	206,072	131,144	663	58.2%	63.6%	131,407	-263
2004	215,694	197,006	125,736	668	58.3%	63.8%	122,349	3,387
2000	202,609	186366	110,826	671	54.7%	59.5%	105,594	5,232
1996	193,651	179,935	105,017	650	54.2%	58.4%	96,390	8,627

[1] Source: Current Population Survey

[2] Source: Office of the Clerk- U.S. House of Representatives

Table 2. Missing Data and Response Patterns in a Hypothetical Population Based on 2008 CPS Voting Responses and Coverage

	Unweighted	Weighted [5]	Weighted and Imputed[6]
Total in-universe population	100,000		
Potential Interviews[1]	87,300		
Population in interviewed households[2]	79,581	100,000	
			100,000
Respondents to voting question[3]	69,506	87,340	
Answered "yes"	51,677	64,936	74,349
Answered "no"	17,829	22,404	25,651
Missing[4]	10,075	12,660	
Percent of total population that voted		65%	74%

[1] Based on .873 coverage ratio

[2] Household response rate of 91%

[3] Number of interviews that completed basic CPS, and did not have No response, Don't Know, or Refusal for the voting question.

[4] Includes "Don't Know," "Refused," and "No Response"

[5] Average Weight calculated as the ratio of interviews over total poption: 100,000/79,581

[6] All missing data assigned to "yes" or "no" categories at the same rate as those who responded.

Table 3. CPS Over-Report of Voting by State			
State	**2000**	**2004**	**2008**
District of Columbia	32.2%	18.6%	15.1%
South Carolina	24.6%	17.4%	9.3%
Mississippi	22.0%	10.8%	11.6%
Alabama	17.2%	9.4%	1.2%
Indiana	16.6%	5.3%	0.3%
Oklahoma	15.9%	5.3%	3.0%
Louisiana	15.0%	6.4%	9.6%
West Virginia	12.9%	5.6%	3.9%
Missouri	12.7%	3.1%	-2.7%
Georgia	9.4%	0.9%	6.6%
Texas	9.3%	7.3%	4.4%
North Dakota	8.6%	5.5%	1.4%
Virginia	8.1%	-1.9%	-2.0%
New Mexico	8.1%	10.7%	1.9%
Utah	7.6%	10.1%	-1.4%
Maryland	7.5%	1.2%	-0.8%
Delaware	7.5%	2.6%	-1.1%
Arizona	7.3%	11.2%	8.9%
Kansas	7.1%	0.0%	-1.4%
Rhode Island	7.1%	6.8%	7.5%
Kentucky	6.5%	7.5%	6.9%
Illinois	6.1%	7.5%	-1.6%
New Jersey	5.9%	2.3%	-6.0%
Nevada	5.2%	5.0%	6.1%
Tennessee	5.1%	-4.9%	-3.2%
California	4.8%	3.1%	2.0%
Maine	3.9%	-0.6%	-2.1%
North Carolina	2.7%	3.9%	1.7%
Michigan	2.6%	-0.4%	-2.7%
Ohio	2.6%	-2.5%	-3.9%
Wyoming	2.5%	1.3%	-1.9%
Nebraska18	1.9%	1.9%	5.3%
Washington	1.6%	-0.3%	1.2%
Pennsylvania	1.5%	1.3%	-4.4%
Arkansas	1.5%	8.1%	0.5%
Massachusetts	1.4%	5.4%	-1.9%
Wisconsin	1.3%	0.4%	-3.2%
Florida	0.7%	-3.1%	-5.2%
New York	0.6%	3.4%	-2.1%
New Hampshire	0.3%	-0.2%	-0.4%
Iowa	0.0%	1.0%	-2.4%
Idaho	-0.3%	-2.2%	-1.7%
Oregon	-0.3%	4.7%	-0.5%
Montana	-0.5%	7.0%	-3.5%
Vermont	-1.5%	1.2%	-5.2%
South Dakota	-1.7%	-2.6%	2.1%
Minnesota	-2.6%	2.1%	-5.2%
Alaska	-5.4%	-6.3%	-6.8%
Colorado	-6.2%	-1.5%	-3.9%
Hawaii	-7.6%	0.9%	0.2%
Connecticut	-8.7%	-3.5%	-2.2%
Total	5.0%	2.8%	-0.2%

Source: Current Population Survey
Source: Office of the Clerk- U.S. House of Representatives
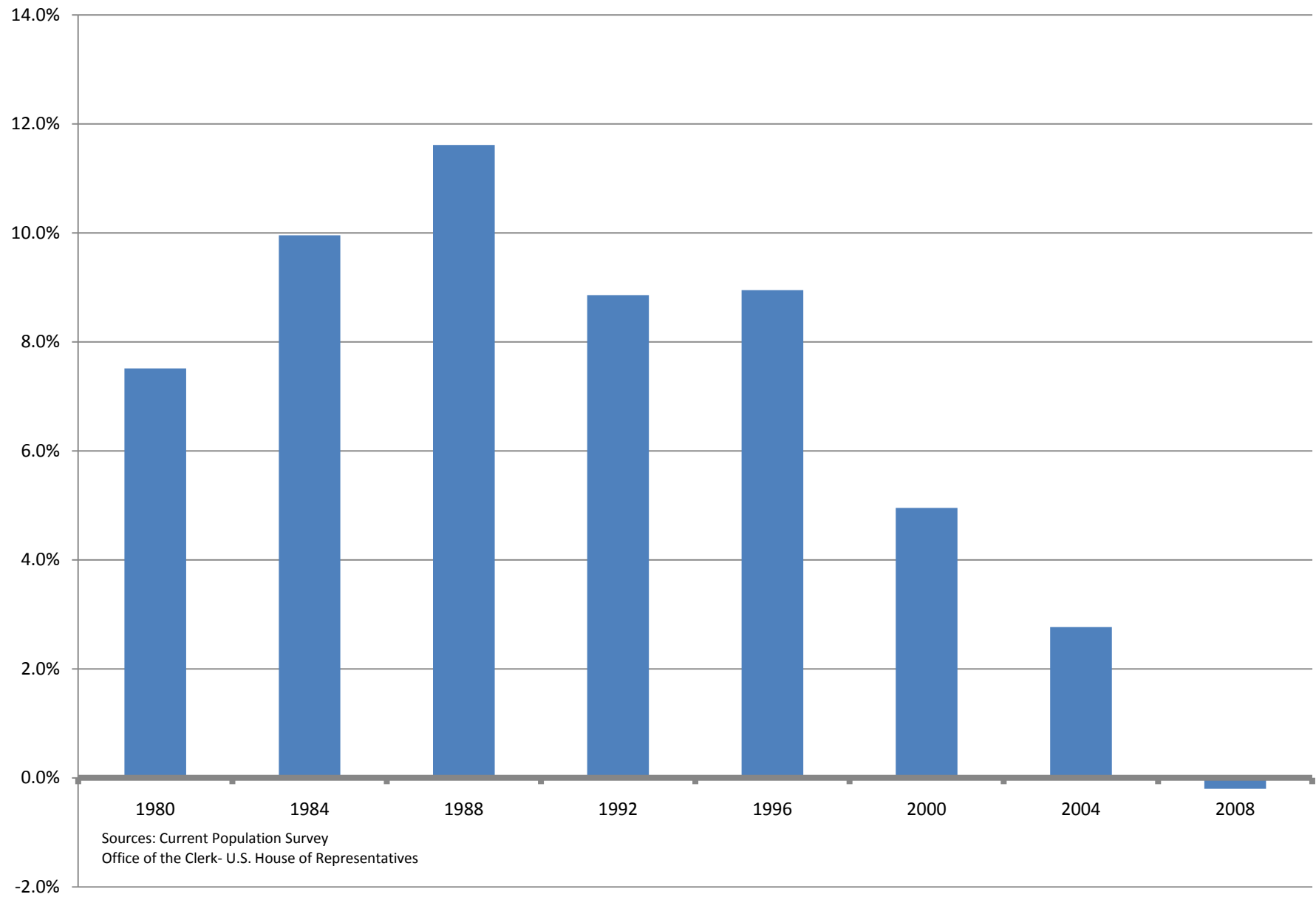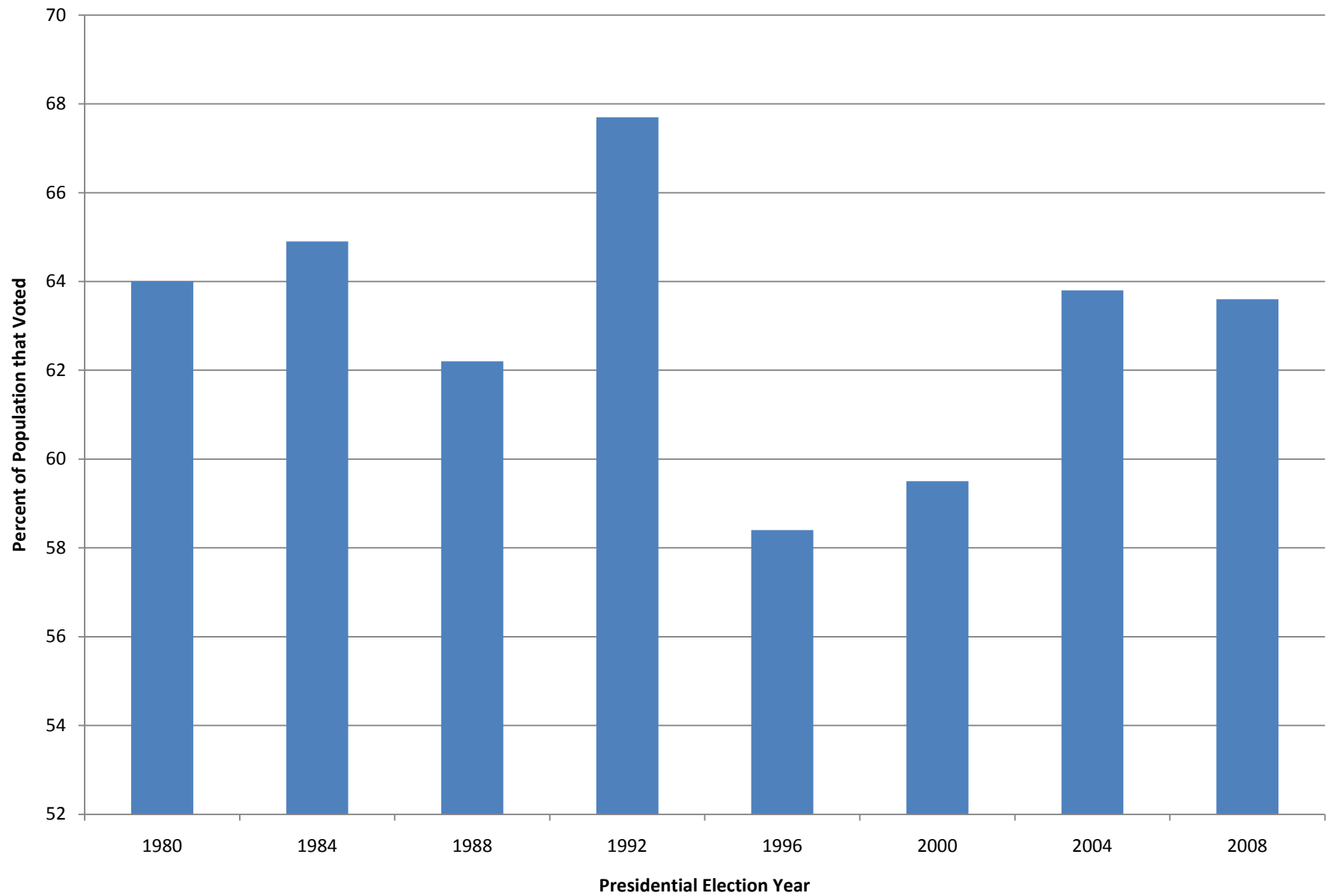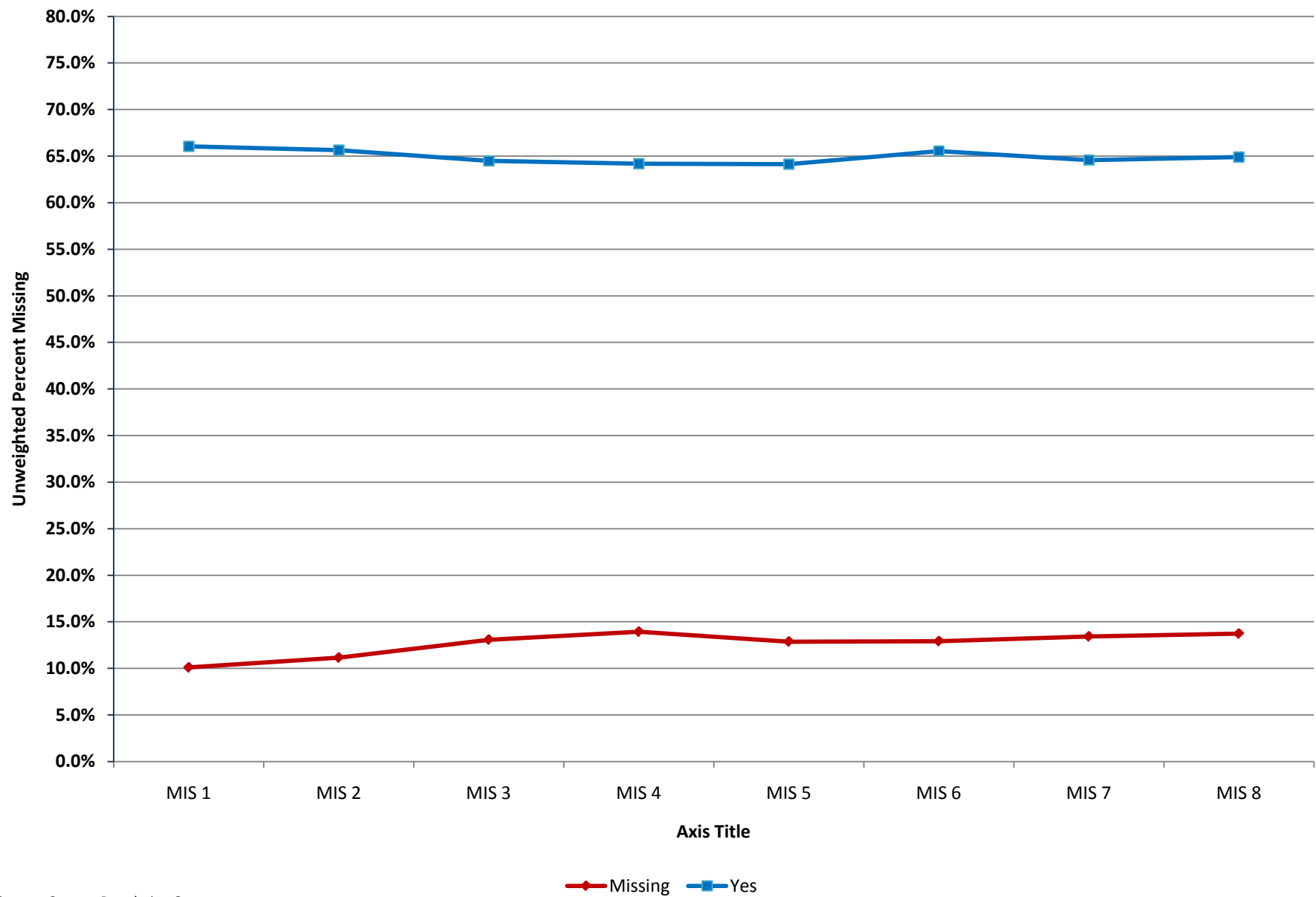
Figure 1. Voting Turnout Overreport from 1980 to 2008

Sources: Current Population Survey
Office of the Clerk- U.S. House of Representatives

Figure 2. Presidential Turnout 1980 to 2008

Percent of Population that Voted

Presidential Election Year
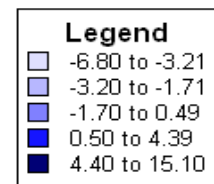
Source: Current Population Survey

Figure 3. Month in Sample by Voting Report 2008

Source: Current Population Survey

Figure 4. CPS Overreport of Voting by State

Legend
- ☐ -6.80 to -3.21
- ☐ -3.20 to -1.71
- ☐ -1.70 to 0.49
- ■ 0.50 to 4.39
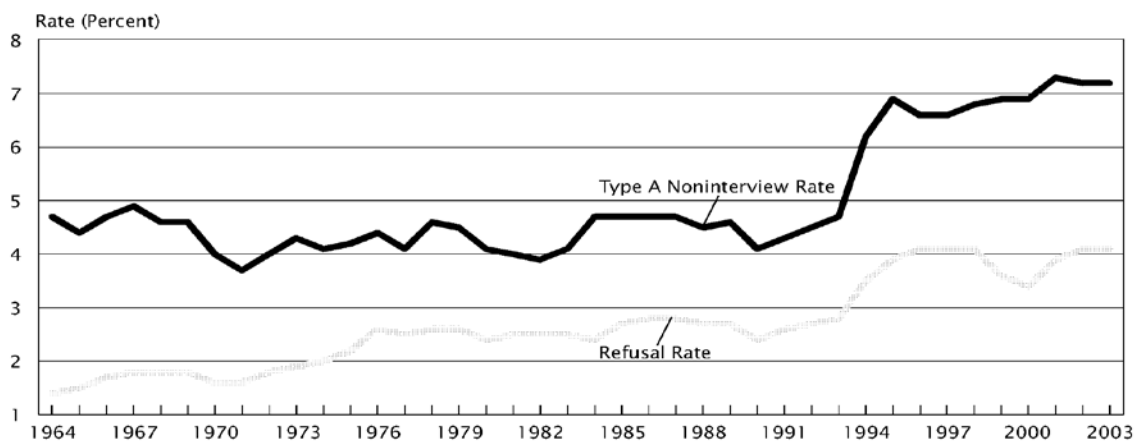- ■ 4.40 to 15.10

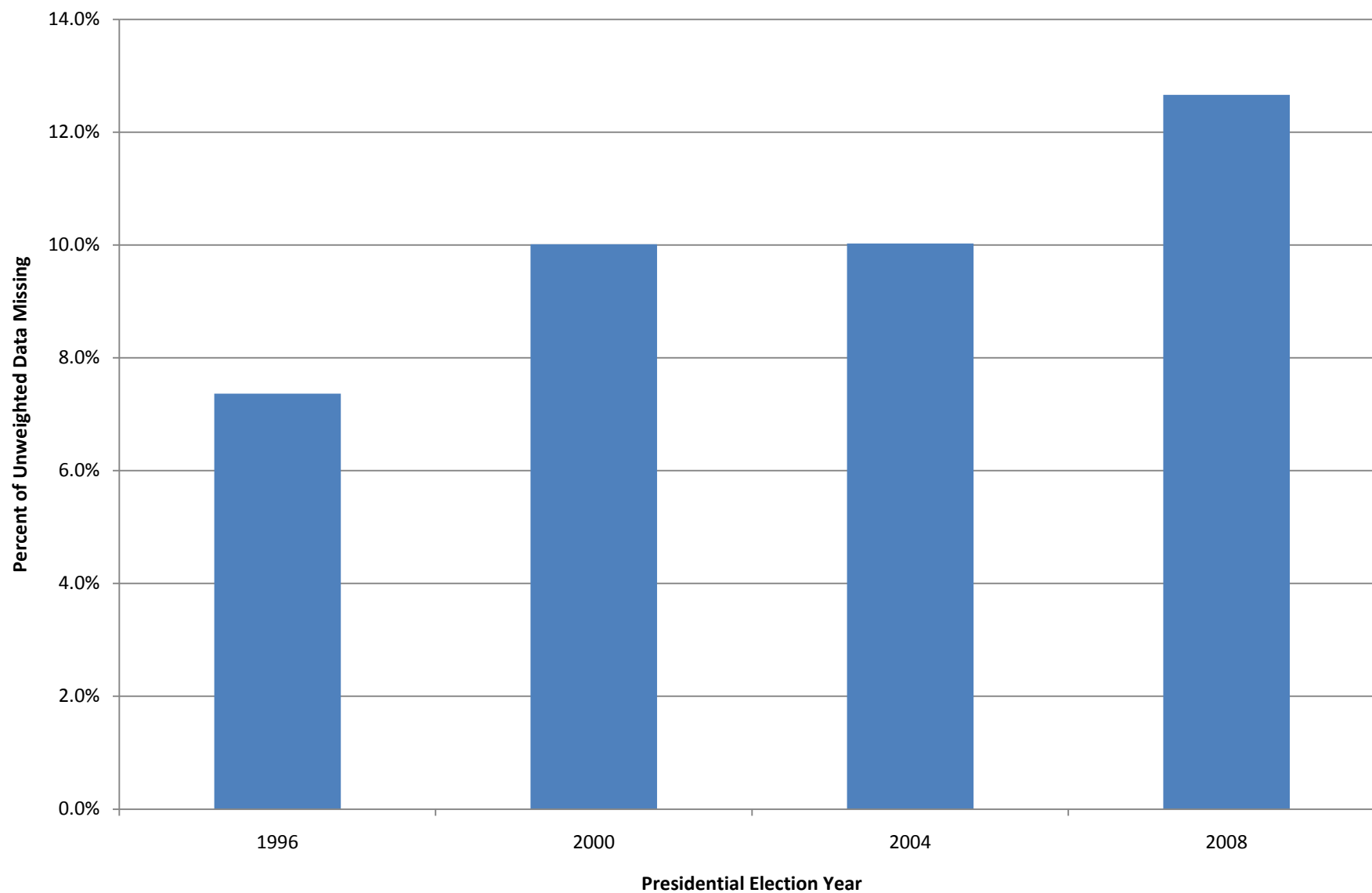Source: Current Population Survey 2008

Figure 5.

Average Yearly Type A Noninterview and Refusal Rates for the CPS 1964–2003, National Estimates

Souce: CPS Survey Design and Methodology Technical Paper 66

Figure 6. Missing Data on Voting Responses by Presidential Election 1996 to 2008

Source: Current Population Survey

Appendix Table 1. SAS Output from Simulated Imputation

Frequency Percent Row Pct Col Pct	Table of CPS Variable by Simulated Imputed Vote Variable		
	CPS Variable	Simulated Imputed Vote Variable		
		yes	no	Total
	DK, R, NR	20,136,835	8,244,215	28,381,050
		9.77	4	13.77
		70.95	29.05	
		13.31	15.05	
	yes	131,143,947	0	131,143,947
		63.64	0	63.64
		100	0	
		86.69	0	
	no	0	46,547,416	46,547,416
		0	22.59	22.59
		0	100	
		0	84.95	
	Total	151,280,782	54,791,631	206,072,413
		73.41	26.59	100